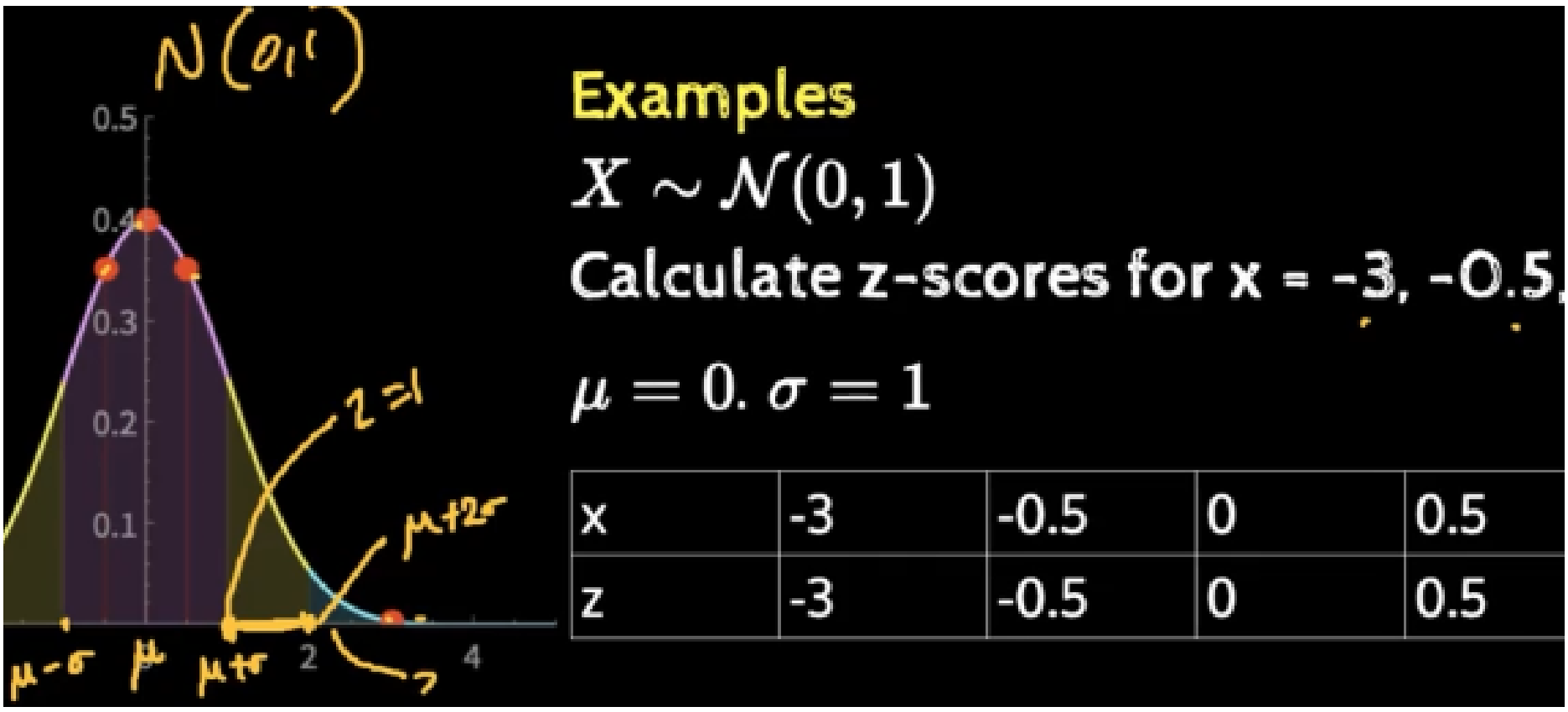


x	-3	-0.5	0	0.5
z	-3	-0.5	0	0.5

Week 19 : Central Limit Theorem

pending tasks

type

Central Limit Theorem

def CLT - (t = 6:03)

- Though the sample parameters of mean and variance can be deduced from the population parameters, nothing more can be said about the distribution per se .i.e. the distribution shape of the sample mean.
- The sample mean $\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$ is sum of n independent and identically distributed (iid) random variables.
- Central Limit Theorem (CLT) is useful when one random variable is a sum multiple independent random variables.
- **CLT : If $X_1, X_2, ...X_n$ are random samples from a population with mean μ and finite deviation σ, then the sum $X_1 + X_2 + ... + X_n$ will converge to $N(n\mu, \sigma\sqrt{n})$ in the limiting condition $n \to \infty$.**
- The result $E[\bar{X}] = \mu$ and $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ remain true for large n. This is also satisfied by the CLT.
- The key insight from CLT is that **independent of the PDF of the population,** the PDF of sum of many samples tends to the **normal distribution.**

Demo 01

- Five distributions ,namely normal, bimodal, Gumbel, Uniform and Weibull were taken, all with mean 0.
- For each distribution it was observed that the distribution of the sum of 'n' values for 'k' samples tended towards a normal distribution with increasing 'n' and 'k' , irrespective of the distribution.
- CLT only relates the distribution of the sum with the normal distribution, the number of samples required for this is experimental.

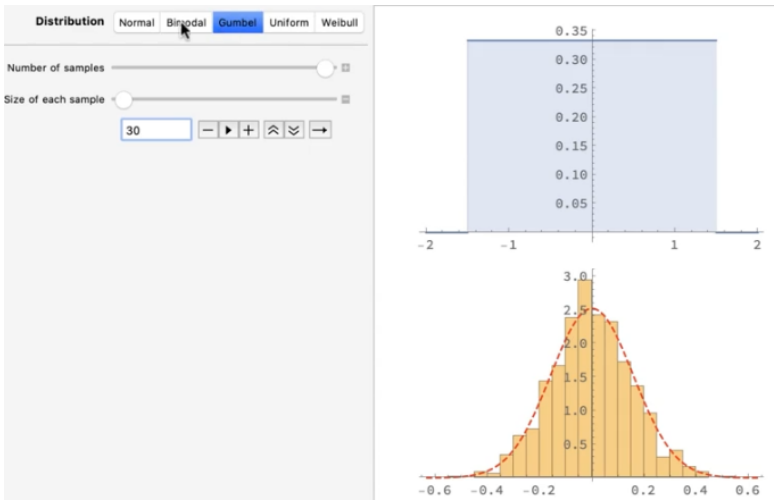


Fig.1 Demo for CLT using uniform distribution.

- The sample size of 30 generally gives a good approximation of the normal distribution.

Alternative Version of CLT

def n^{th} moment - (t = 7:45)

- Instead of sum of sample elements, the sum of shifted and scaled version of the sample elements is taken.
- **Alternate version of CLT : If $X_1, X_2, ...X_n$ are random samples from a population with mean μ and finite deviation σ, then the sum $\frac{X_1-\mu}{\sigma} + \frac{X_2-\mu}{\sigma} + ... + \frac{X_n-\mu}{\sigma}$ will have a PDF that converges to $N(0,1)$ in the limiting condition $n \to \infty$.**
- To prove the above theorem, two PDFs have to be compared and the convergence of $n \to \infty$ has to be tested.
- Equivalence of μ and σ is a necessary but to not sufficient condition to prove that two PDFs are equal.

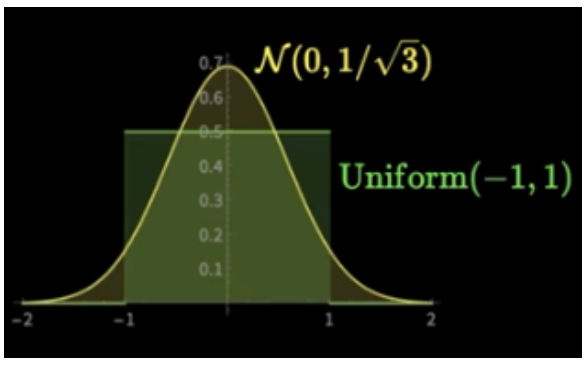


Fig.2 Example of two different distributions with same mean and stddev.

- To prove the equivalence of two PDFs their moments have to be compared. μ and σ are special cases of moments.
- The n^{th} moment is the expected value of the n^{th} power of the samples. It is given by

$$E[X^n] = \sum_i x_i^n p(x_i),\ discrete$$
$$= \int_{-\infty}^{+\infty} x^n f(x) dx,\ continuous$$

- The zeroth moment $E[X^0] = 1$, the first moment $E[X^1] = \mu$ and the second moment of mean adjusted variable $X = X' - \mu$, $E[(X' - \mu)^2] = \sigma^2$ for all distributions. Thus, the axioms of probability, the measures of centrality and spread are moments of the distribution.
- The third moment with shifted and scaled variable $X = \frac{X'-\mu}{\sigma}$, $E[(\frac{X'-\mu}{\sigma})^3]$ gives the measure skewness of the distribution.

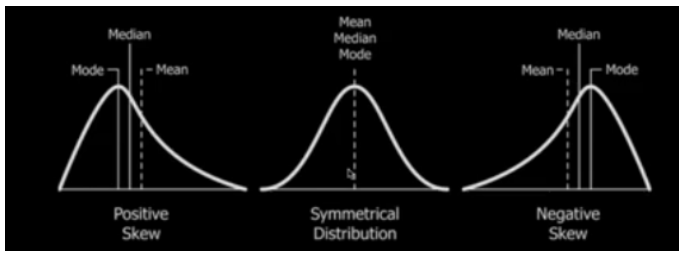


Fig.3 Skewness in distributions.

- Positive skewed distribution is right screwed, where mean> median and mode and the opposite is true for a negative skew. A symmetrical distribution has a skew of 0.
- The moments are essentially signatures of the distribution. Thus, two PDFs can be compared by matching their moments.

CLT - Attempt at Proof

- How to test convergence of $n \to \infty$? What must be the property of the PDF that satisfies a convergence condition?
- In alternative CLT every distribution converges to the standard normal distribution. So what must be the properties satisfied by the standard normal distribution?
- Let $Y_k = X_1 + X_2 + ... + X_k$
  if $Y_n \sim N(\mu, \sigma)$, then CLT requires that $Y_{2n} \sim N(\mu', \sigma')$
- If Y_{2n} is to be normally distributed, then the sum of two normal distributions, (Y_n, Y_n) should also be normally distributed.
- Therefore, it is to be inspected if the sum of two normally distributed RVs is also normally distributed. This turns out to be true.
- For two RVs $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$
  then, $X + Y \sim N(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$
- The normal distribution is an **attractor.** Once the distribution becomes normal it remains normal.

// two PDFs can be compared using Moments. Sum of normals is also a normal distribution.

Implications of CLT

- Independent of the PDF of population, PDF of $\bar{X}$ is normal, under the assumption of CLT that sample size n is large.
- From $\mu \; and \; \sigma$, the likelihood of getting a value greater than $\bar{X}$ can be computed.

Likelihood of Sample Mean

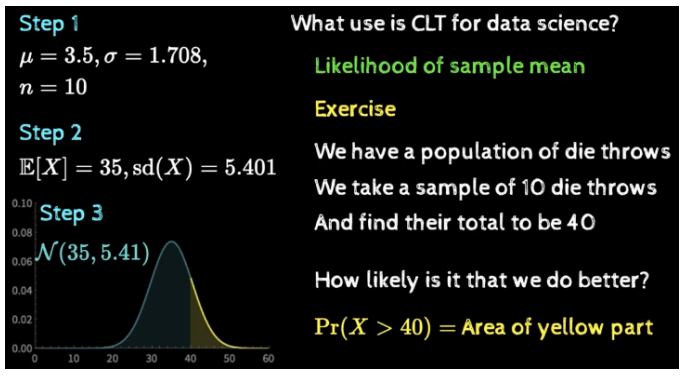


Fig.4 An example for calculation of likelihood .

- To find the solution, since the distribution is continuous $f(x)$is integrated from 40 to ∞ giving the answer 0.177.
- Since this is a recurring calculation in problems using the normal distribution, better ways have been devised to compute the area under the curve.

Computing Area Under N

- The area under the normal distribution $N(\mu, \sigma)$ is mapped to the standard normal distribution $N(0, 1)$ using the **z-score.** $Z = \frac{X - \mu}{\sigma}$**,** where $Z \sim N(0, 1)$.
- z-score is a scalar to scalar mapping. It denotes the number of standard deviations a value is away from the mean. This z-score holds for all distributions, not only for the normal distribution.

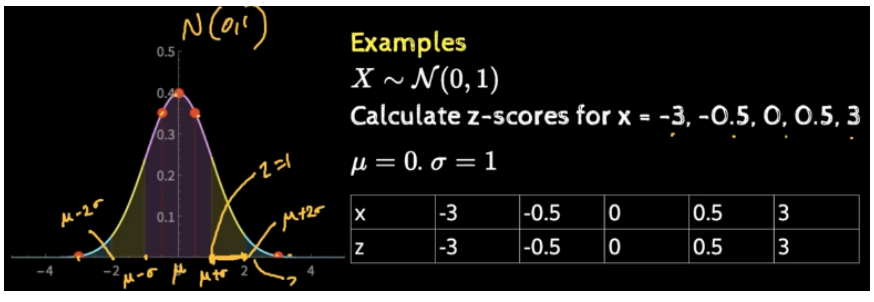


Fig.5a An example of z-score calculation.

- Larger magnitude of z-score implies that the probability of the value is low in the normal distribution. The sign of the z-score indicates which side of the mean the value lies.

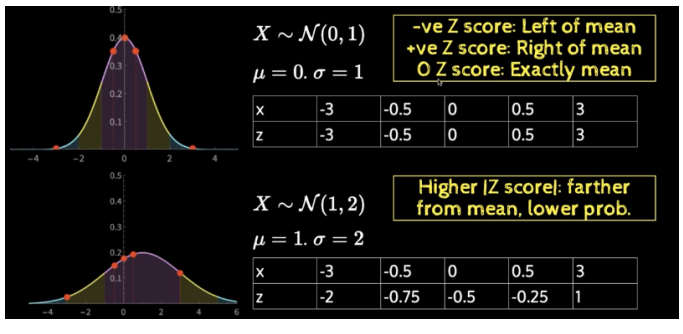


Fig.5b insights on z-score.

- Thus, to calculate given area under the normal distribution can be mapped to area under the normal distribution as follows:

$$Pr(X < a) = Pr(\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma})$$
$$= Pr(Z < \frac{a-\mu}{\sigma}))$$

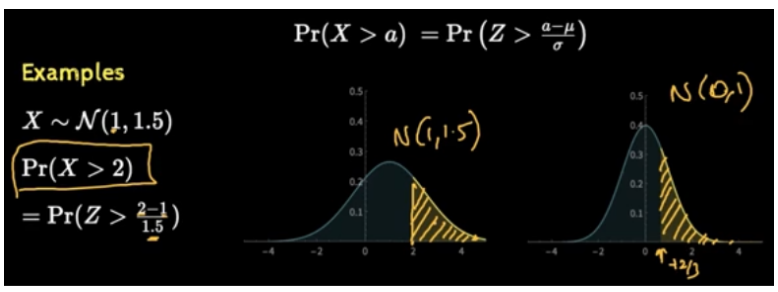


Fig.5c Mapping area under normal distribution to area under standard normal distribution

Demo 02

- Except the computation of z-score, the calculation of area is independent of $\mu\ and\ \sigma$.

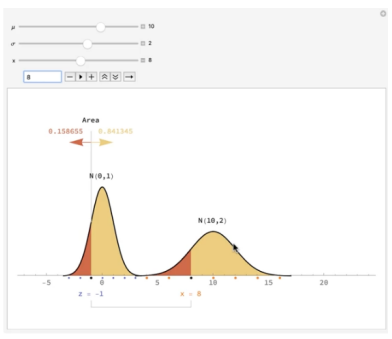


Fig.6 The mapping of PDF to normal distribution.

Special Significance for N

// z-score helps to calculate areas under the curve, thus, probabilities of the kind Pr(X>a), where X is distributed as per some normal distribution by mapping it to the standard normal distribution.

- $PDF \rightarrow_{(CLT)} normal\ dist. \rightarrow_{(z-score)} std.\ normal\ dist.$
- Initially look-up table was devised to improve the efficiency of the area computation.
- **Some relations on the area**

$$Pr(Z > a) + Pr(z \leq a) = 1$$
$$Pr(Z > a) = Pr(z < -a)$$

$Pr(z < -3.5) \approx 0$

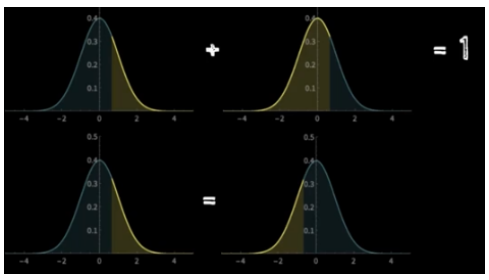


Fig.7 relations 1 and 2.

- It can be inferred from the above relations that it is sufficient to store Pr(Z<a) for $a \in [-3.5, 0]$.
- The standard normal distribution table is used to find the probability of a given z-score.
- Pr(a < X < b) = Pr($\frac{a-\mu}{\sigma}$ < Z < $\frac{b-\mu}{\sigma}$) , this requires two look-ups.

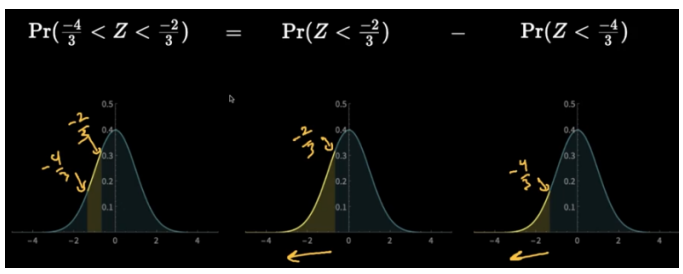


Fig.8 Example to compute Pr(a < X < b)

Likelihood of Sample Mean

- Solving the **Likelihood of Sample Mean** example considered earlier, using CLT:

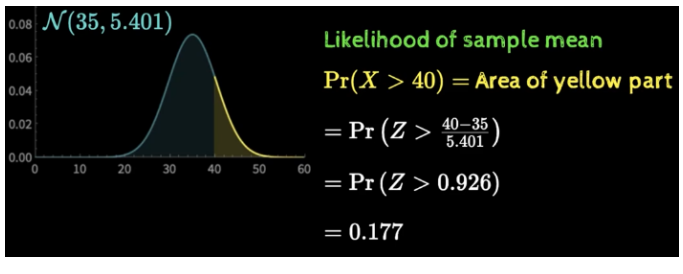


Fig.9 Calculating area under curve using CLT.

- The probability computation without CLT involves plotting the multinominal distribution of $f(x)$for all possible values, ranging from 10 to 60 and adding the required values. Pr(Q > 40) = f(41)+f(42)+....+f(60).
- The answer is 0.157, which is shows that the calculation using CLT, 0.177 was an over-estimation. If the CLT method is to be used often, then this issue of over-estimation has to be addressed.

Super-Impose N

def continuity correction - (t = 11:03)

- The CLT only insures that the two PDFs are approximately the same. The computation of area is an added improvisation not insured by the CLT.
- The resultant over-estimation is from the exclusion of f(40) in the original computation (discrete multinomial plot) , values only from f(41) are included. Where as in the CLT(continuous curve) method, values greater than 40 have been considered. This is shown in Fig.10a.

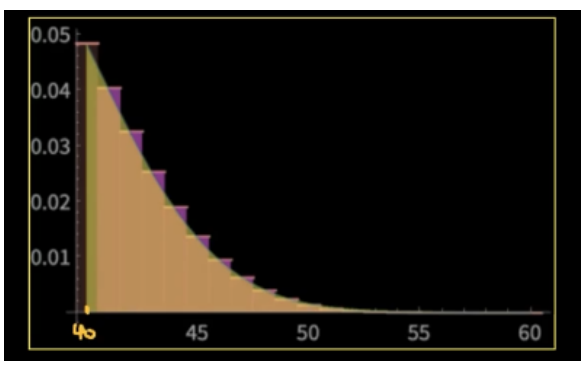


Fig.10a Area under curve vs area covered by the bars., an over-estimation.

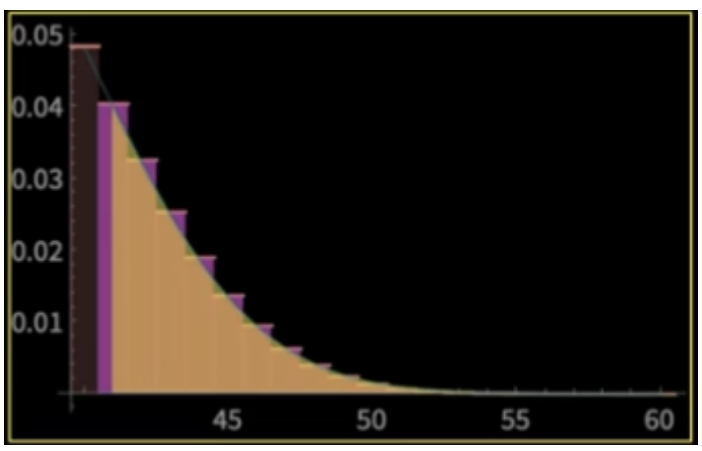


Fig.10b Area under curve vs area covered by the bars., an under-estimation.

- If Pr(X>41) is taken, then the result using z-score is 0.133. This is an under-estimation. This is shown in Fig.10b. As seen this is caused by the missed area under the curve which is covered in the bar.
- Thus, a values in the middle is taken , Pr(X > 40.5) which results in 0.156 and this is a fairly better approximation. This is called as **continuity correction,** used while approximating a discrete distribution with a continuous distribution to get accurate estimates.

Approximating Distributions

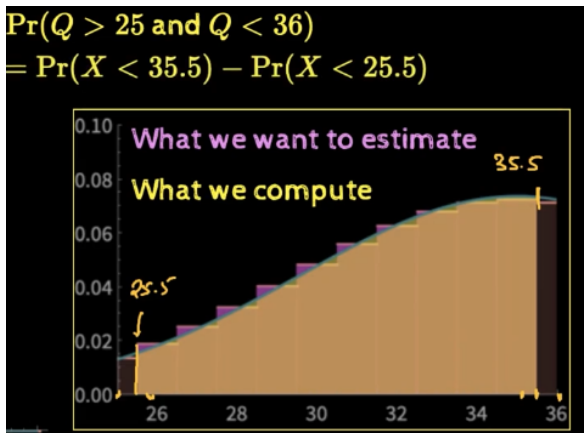


Fig.11a Another example of using continuity correction.

- CLT is a good approximation of computing likelihood of values for larger values on n.
- CLT can also be used to approximate distributions.

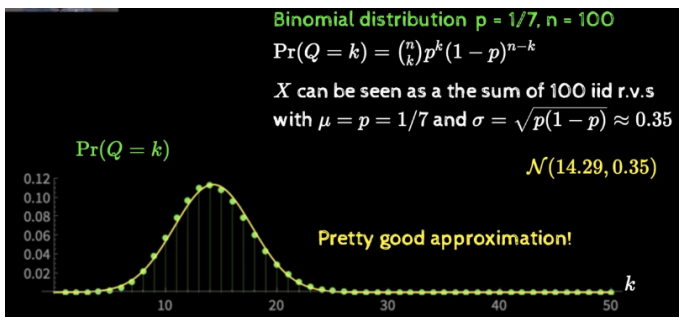
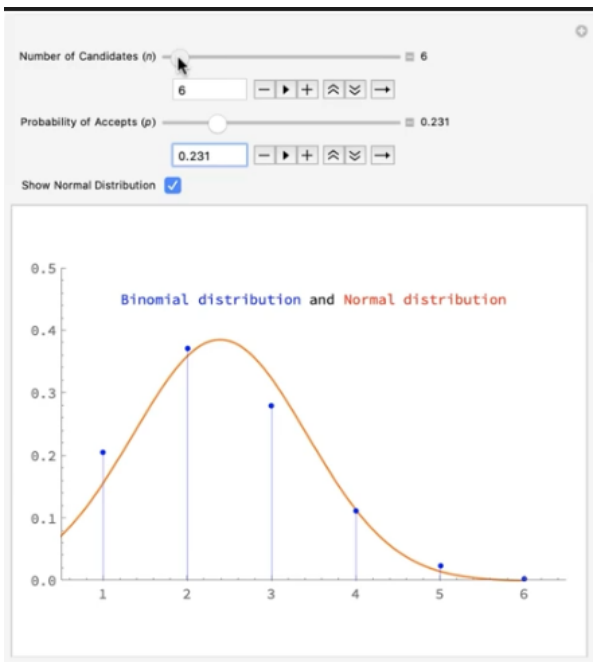


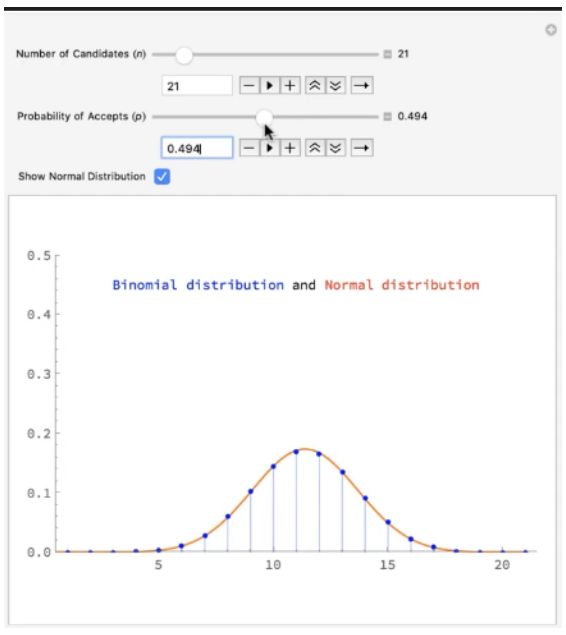
Fig.11b An example of approximating binomial distribution to normal distribution using CLT.

- In the example shown in Fig.11b, the expensive computation of probability can be replaced by using CLT for approximating the PMF to PDF of normal distribution, thus, making calculations easy.

Demo 03

- Values of p (probability of success) and n (sample size) under which the approximation to normal distribution holds better for a binomial distribution .
- It is observed that, in the extreme values of p where the distribution is not symmetric the approximation to normal distribution using CLT is not so good. The approximations improve with higher values of n.





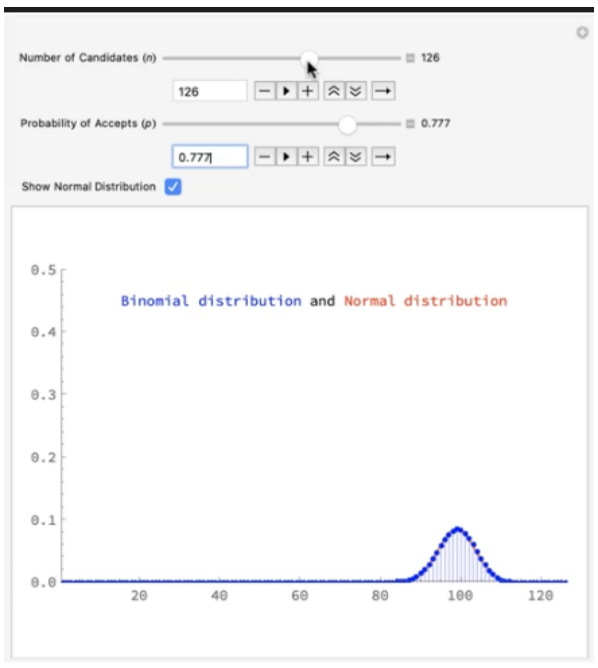


Fig.12c p close to 0.5 giving a better approximation.

Fig.12a Extreme p and small n, a bad approximation.

Fig.12b Extreme p and large n, obtaining a good approximation.

Normal Approximation of Binomial Distribution

- Normal approximation of binomial distribution holds under the CLT. It was observed that the accuracy of approximation were higher when $p \approx 0.5$ and for larger values of n.
- **Rule of thumb : np > 10** and **n(1 - p) > 10 (**for CLT we had sum of 30 values will ensure that the result is normally distributed).
- Bell curves are observed in a lot of real world data and are very close to being a normal distribution.

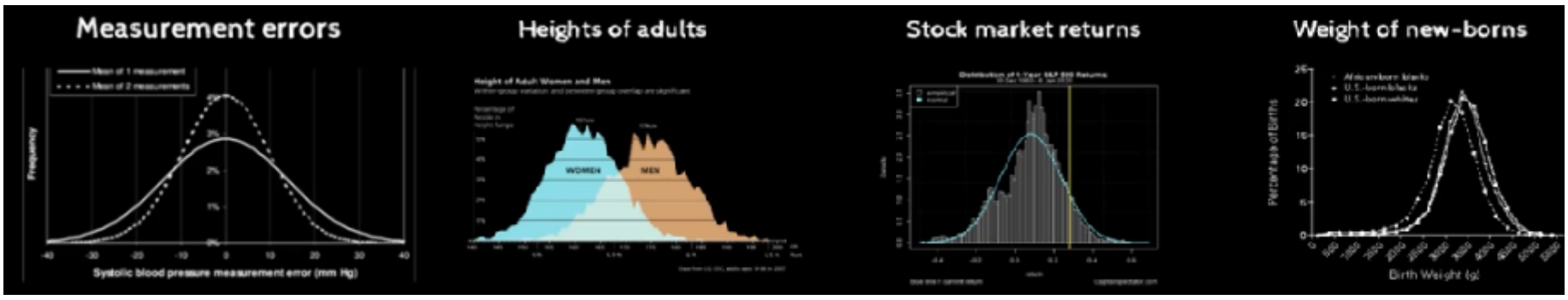


Fig.13 Examples of bell curves observed in real world data.

General CLT (the definition of CLT states same population characteristics for all RVs) : $X_1, X_2 ..., X_n$ with different $\sigma \; and \; \mu$

If $X_1, X_2 ..., X_n$ are independent random variables each with mean μ_i and finite variance σ_i^2, and let $s_n^2 = \sum_i^n \sigma_i^2$, then the sum $\sum_{i=1}^n \frac{X_i - \mu}{s_n}$ converges in distribution to $N(0,1)$ under some conditions (Roughly, no σ_i^2 is comparable to the sum s_n^2).

- The implication of above definition is that if RV X can be written as the sum of many independent RVs, then X is likely to have a bell-shaped distribution.
- As more RVs are added (convoluted) together, the resultant distribution resembles a normal curve more closely. Thus, many real world data have normal distribution as they are a result of convolution of multiple RVs.

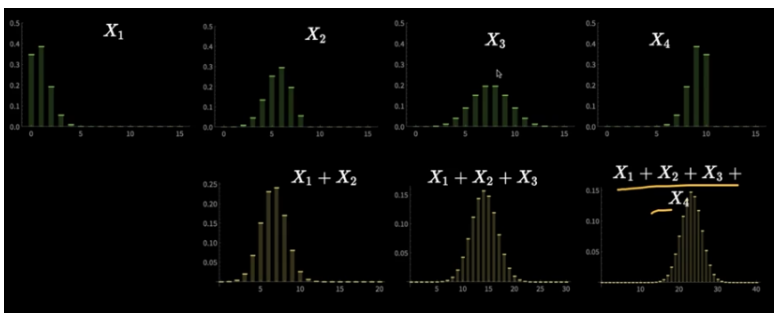


Fig.12 Examples of convolution of RVs.